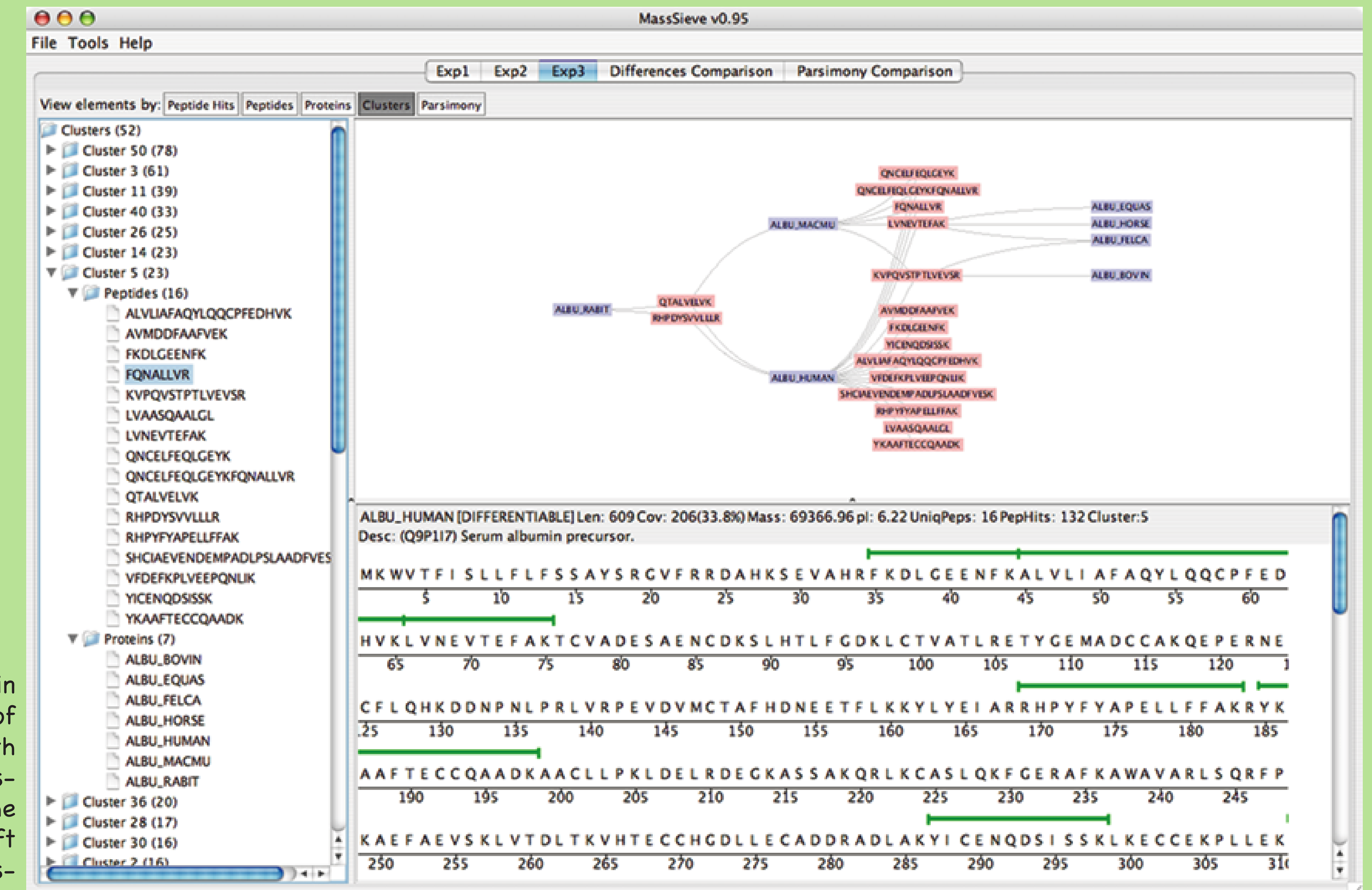
MassSieve: A New Tool for Mass Spectrometry-based Proteomics

Douglas J. Slotta, Melinda A. McFarland, Anthony J. Makusky, Sanford P. Markey
Laboratory of Neurotoxicology, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892

National Institute of Mental Health

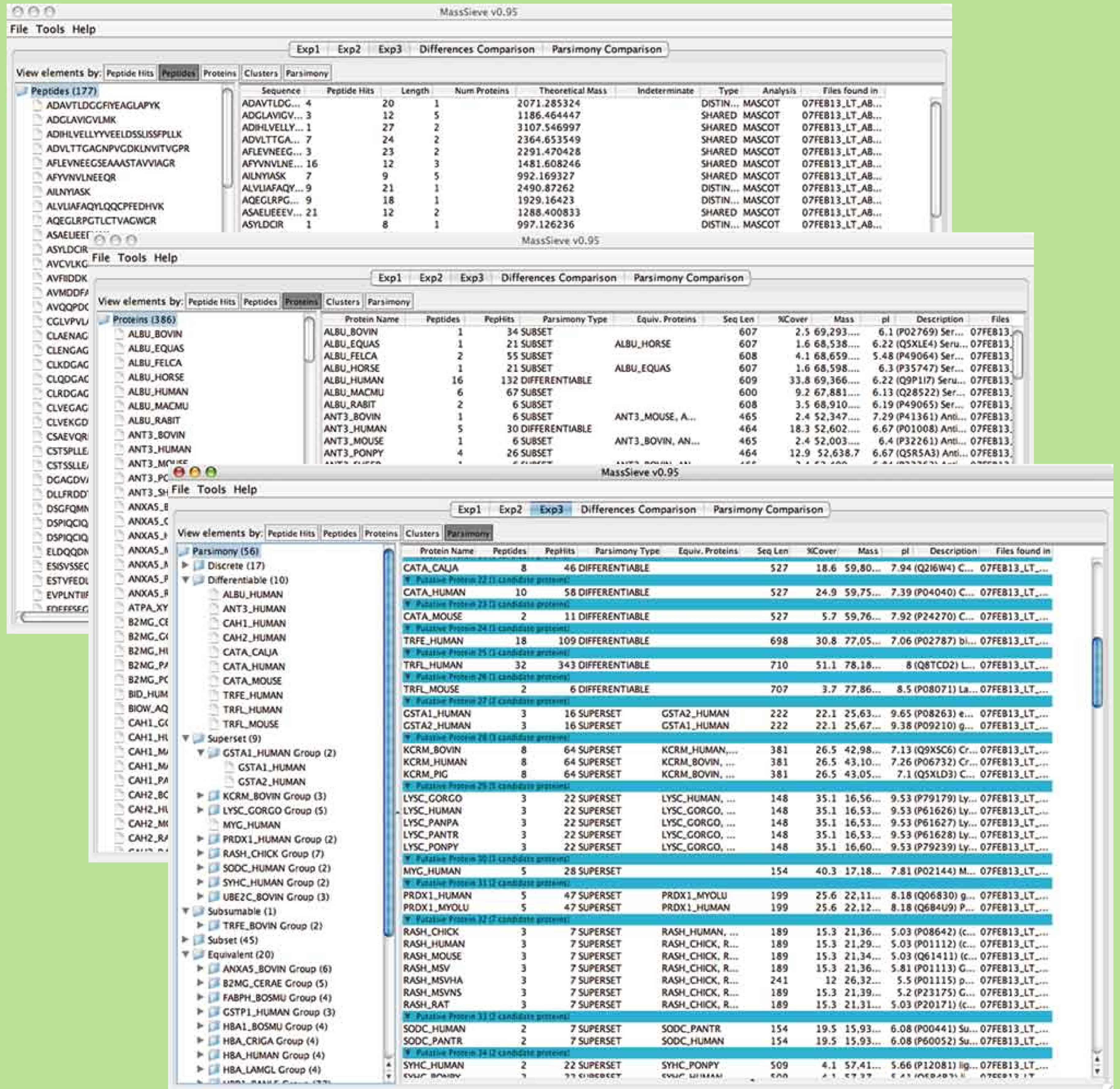
Overview:

The success of peptide sequence assignment algorithms such as OMSSA and Mascot for mass spectrometry has led to the need for a tool to evaluate the results. DBParser is such a software tool previously developed by the LNT lab for this purpose. Its value for parsimonious analysis of proteins associated with experiments has led to its use for analyzing larger datasets than initially anticipated (100's of data files with millions of spectra). MassSieve builds on this experience and is designed as open source protein assignment software that can be scaled to apply parsimony principles to very large experiments without data set size limitations. In addition it allows a more interactive view of the results.



Figures 2, 3, 4:

In addition to the cluster tree shown above in Figure 1, the experimental results can also be viewed as a list of peptides and their associated proteins or as a list of proteins and their associated peptides. The proteins can also be listed in terms of their parsimony categories. Clicking on an protein or peptide will display information about that object in the lower right-hand section. Clicking on a category (e.g. Shared or Superset) will produce a list of the contained objects in the upper right-hand window.



Figures 8 & 9:

A list of all proteins and selected information about them is shown in Figure 7 above. Figure 8 on the right shows the same list for a parsimony comparison with the Unique Peps and the PepHits field expanded. These lists (like all other lists in MassSieve) can be copied and pasted into another program (e.g. Excel) or exported as a comma separated value formatted file.

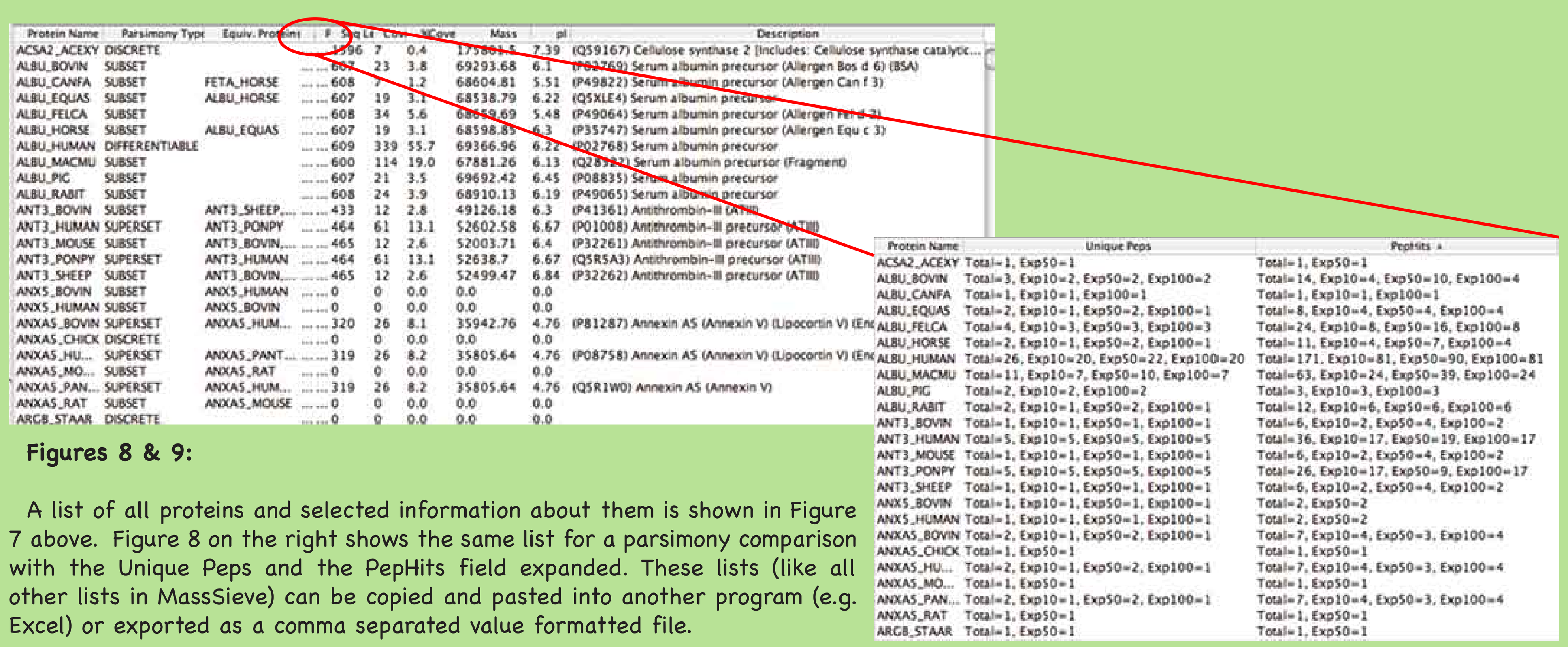


Figure 1:

This is the main window of MassSieve, with three main display areas. The tree on the left shows the clusters of related proteins and peptides, and can be changed to display the trees in Figures 2, 3, and 4. The upper right-hand display area has a graph showing the relationship between the peptides and proteins and can be switched to show the charts in Figures 8, 9, and 10. The lower right-hand area displays a graphical view of a protein sequence. The green bars denote peptides found by the mass spectrometer. This view can be changed to show a list of peptide hits for an individual peptide as shown in Figure 11.

Rules of Parsimony:

Since a given peptide sequence can map to more than one protein, a method is needed to reduce the complexity of the resulting set of proteins. The following parsimony rules and definitions are designed to reduce the number of proteins and yet still account for all of the observed spectra.

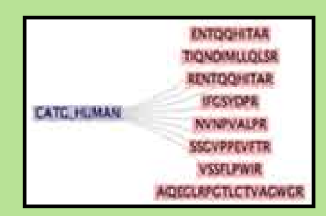


Figure 5:

A graph showing a discrete protein containing only distinct peptides.

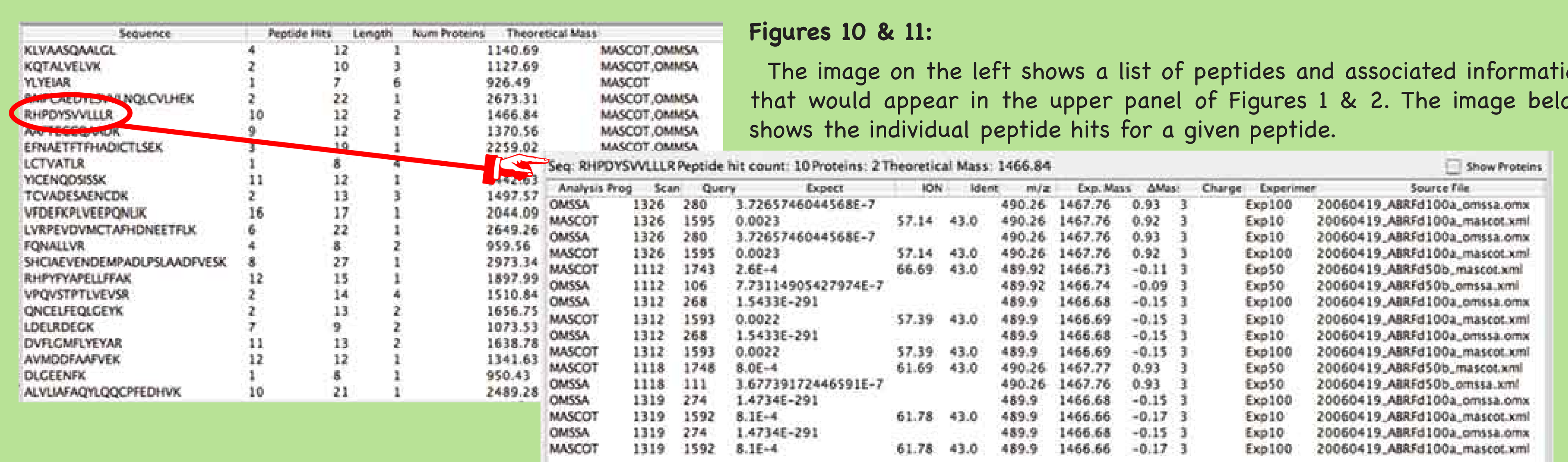
Peptides

- Indeterminate: For a given scan, only the top scoring hit is used. If there is more than one match that ties for the top score, then the peptide is indeterminate.
- Distinct: A peptide that is assigned to exactly one protein.
- Shared: A peptide that is assigned to more than one protein.

Proteins

- Discrete: A protein identification that is identified by only discrete peptide(s), e.g. CATG_HUMAN in Figure 5.
- Differentiable: A protein identification that can be distinguished from other proteins because it has at least one distinct peptide that is not present in other set of peptide(s) and at least one shared peptide that is present in other set of peptide(s), e.g. KCRM_HUMAN in Figure 1.
- Superset: A protein identification contains the shared peptides from at least one other subset protein
- Subsumable: A protein identification contains shared peptides that can be distributed as subsets of two or more other proteins. Formally, subsumable proteins are simply another class of subsets.
- Subset: A protein identification contains peptides common to a larger set of peptides corresponding to another protein identification which is a superset.
- Equivalent: Protein identifications that are based on the same set of shared peptide(s).

All proteins are promoted to their highest category. N.B. a superset, subsumable, or subset protein may still have equivalent proteins.



Figures 10 & 11:

The image on the left shows a list of peptides and associated information that would appear in the upper panel of Figures 1 & 2. The image below shows the individual peptide hits for a given peptide.

Possible uses for MassSieve:

- Merge data from experiments, gel lanes, or fractions into a single dataset.
- Compare and contrast datasets.
- Examine the effects of different filter criteria on a given dataset.
- Use multiple search engines to expand results or produce a more conservative set of results.

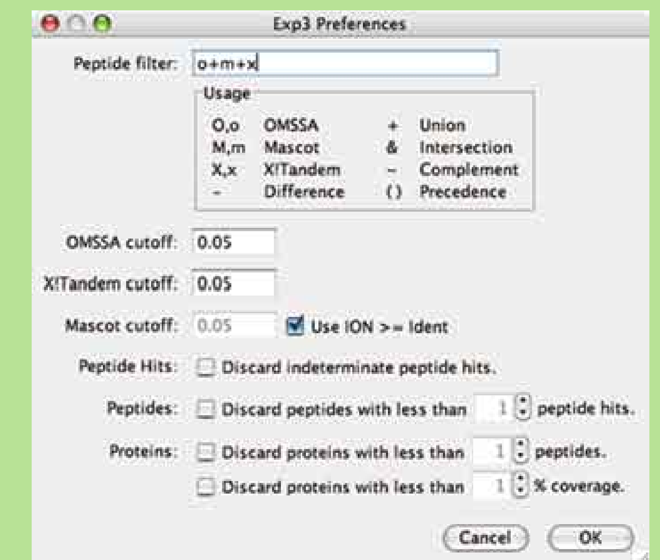


Figure 6:

The peptide results for a given experiment can be filtered based on criteria set by the user. The individual peptide hits are filtered by their score, and the peptides are filtered based on which search engine found them.

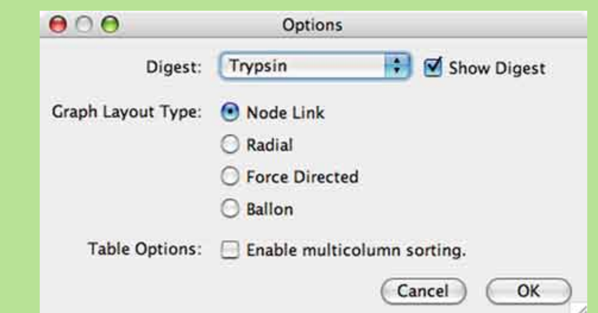


Figure 7:

The options panel allows a theoretical digest to be shown for each protein display and change the graph layout algorithm. In addition, every table in MassSieve is sortable by clicking on the column header, or optionally sortable by multiple columns.

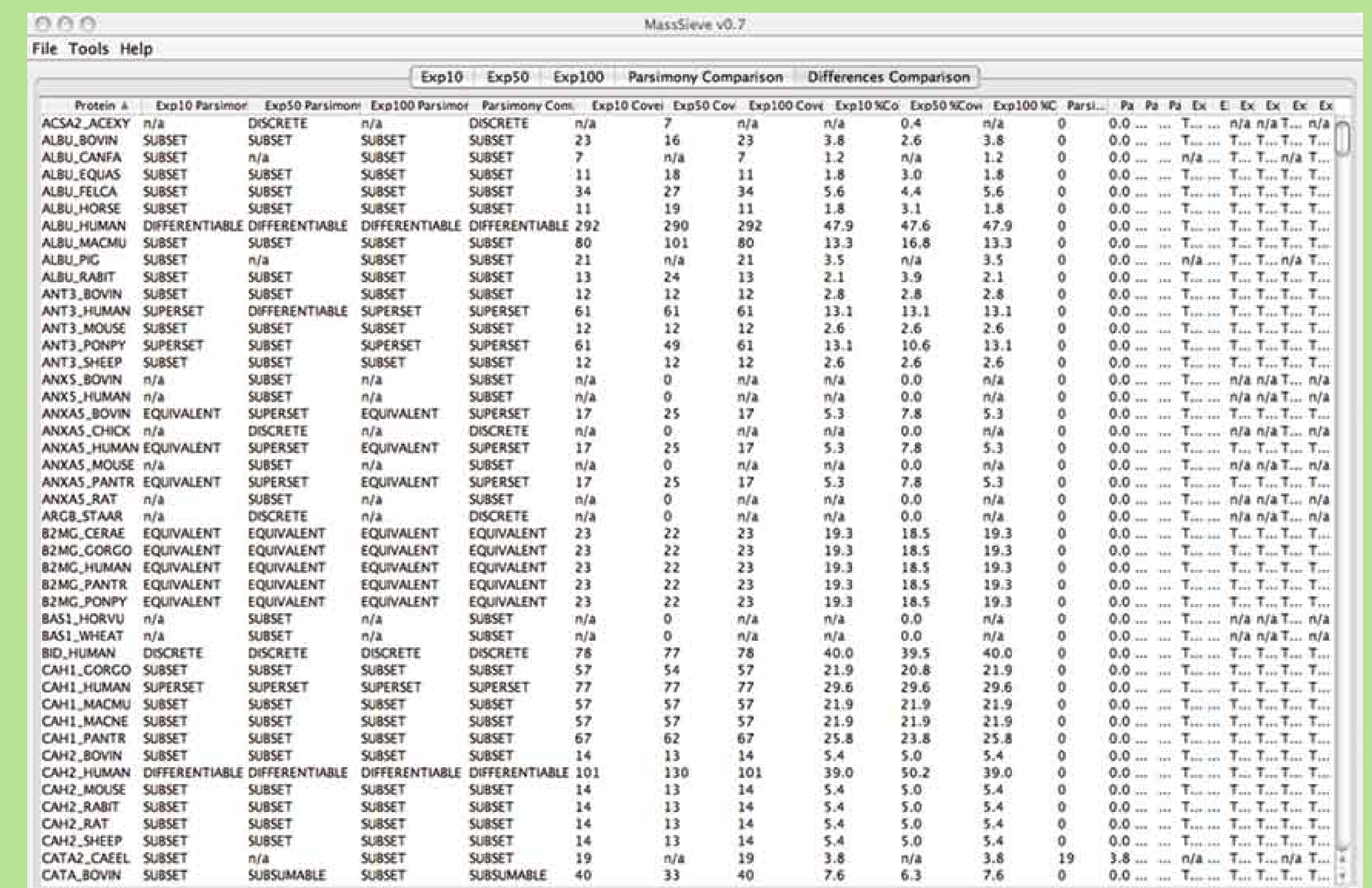


Figure 12:

A tab showing the differences between the experiments for each protein. Note than the parsimony comparison is considered as an experiment for this comparison as well. The recursive nature of this definition allows comparisons ad infinitum.

Future Plans:

- Integrate with a Laboratory Information Management System, such as CPAS.
- Display MS/MS scans for comparisons.
- Add ion current based quantification.
- Support additional search engines.

Acknowledgements:

This research was supported in part by the Intramural Program of National Institutes of Health, NIMH Z01 MH000279.